Text and Data Mining Acquisition and AI Collections Support, Emory Libraries

Prepared by: Chris Palazzolo, Sara Palmer
Feedback and Input from: Humanities, Social Sciences, and Area and International Studies teams; Collections Steering Committee; Scholarly Communications Office
Original Date: 9 July 2018
Reviewed by: Collection Management
Update (for AI): Summer 2024

Purpose of this Document

This document outlines policies for the general acquisition, use, and management of text and data mining (TDM) and/or AI learning corpora, content, and tools at Emory for research and teaching. TDM is here defined as a means to download, extract, analyze, classify, model, or index information using computational tools, algorithms, machine learning, artificial intelligence, or automated techniques. TDM may include the use of AI tools, but not always.

Parameters & Considerations for Acquiring Content:

The following research dataset considerations apply to requests for acquiring corpora that can be used for TDM and AI training (e.g., large language models).

- ❖ **Availability of data:** The Libraries will be more likely to purchase datasets if they are available to more than one researcher. When data can only be available for usage by specific individuals or departments, purchases should be funded through the resources of those parties.
- ❖ **Funding:** The Libraries encourage the use of research and grant funds to share costs of datasets of potential use to the entire Emory community, as well as allocated subject and central library funds
- ❖ **Retention of data:** If retention or destruction of data is a requirement (note that this is unenforceable by Emory), we can only make reasonable efforts to inform users of this stipulation.
- ❖ **Data sharing:** Ideally, data should be made available to co-researchers at other institutions, or at the very least, researchers should be allowed to share their methodology.
- ❖ **Confidentiality:** Does the vendor/provider require any user of the data to sign a confidentiality agreement or anything similar make use of the data? Are there any other sorts of administrative overhead burdens associated with the resource?

Data availability policies of some journals now require authors of submitted manuscripts to provide not just the dataset that was actually analyzed in the manuscript but also the source data that were cleaned and assembled into that analyzed dataset. Journals will make exceptions for data that cannot be shared for terms-of-use reasons, so it will be helpful to have this matter clarified so that the information can be shared with people who wish to use the data.

Other Guidelines Specific to Acquiring, Licensing, and Using TDM and AI training data:

- ❖ Emory prefers that all contracts include explicit clauses for the allowance of TDM and AI training that should not be more restrictive than fair use. Language should be included describing how underlying data should be accessed. Subject librarians will work with Electronic and Continuing Resources (ECR), the Scholarly Communications Office, and/or Collections to determine the feasibility of using content from our licensed resources for such purposes and under what circumstances the content can be used.
- ❖ Where required, updated licenses will be requested by collection management. As much advance notice from the researcher(s) as possible is requested.
- ❖ Most, if not all, licenses restrict automated scraping methods for pulling data. Vendors are concerned about the security of copyrighted data.
    - o Some vendors may require walled garden approaches or other secure methods to control and manage access to large amounts of their data for TDM or large language model training. The Libraries will inform the researcher of these requirements and ask that they provide assurance that the data is managed in a secure location.
    - o Some vendors require the use of a registered API which does not require any additional funds, but which may need some intervention on the part of the subject librarian or head of collections.
- ❖ Generally, we also do not provide licensing or funding for individual TDM projects or AI training data with needs not covered by university wide licenses. However, the Libraries will make efforts to incorporate TDM into existing contracts and/or liaise with the vendor to find an amenable option for the researcher.
- ❖ It is not recommended to use **third-party** AI tools as (1) their use is restricted in our licenses, as they are not secure and (2) many commercial LLM providers may repurpose user data for training. It is advisable for researchers to opt out of any data-sharing.
- ❖ When researchers share data from TDM or AI learning models, it should suffice to simply describe their data sources and methodology rather than providing access to the underlying data or text, as this would be a license violation. Most publishers make exceptions for terms-of-use restrictions related to data sharing.

Purchases of AI Tools by Emory Libraries

AI Tools for purchase or subscription by Emory Libraries must be vetted closely for the following:

Additional price/outlay: AI tools on existing products should not be considered for purchase if they constitute a significant increase in library cost.
Content available: Are there exceptions as to what content from the provider is included for use in the AI tool?
Privacy and re-use: Can the Libraries be assured that any searches and content used in the AI tool will not be re-used in a publicly available AI tool?
Intended use: Is the AI tool primarily for use as a research tool or an instructional tool? If the latter, can it be integrated into an LMS?

Services Provided by Emory Libraries

- Woodruff Library works closely with the Emory Center for Digital Scholarship (ECDS) to provide access to corpora that can be used in TDM.
- Subject librarians maintain, along with ECDS, a libguide that outlines existing, free corpora for manipulation, along with information regarding library-purchased/licensed options and/or restrictions. Researchers should contact their subject librarian if they have any need for AI training or TDM datasets.
- In many cases, TDM corpora will be retrievable through APIs provided by the vendor. These APIs may have stipulations. As these corpora can be relatively large, it is up to the researcher to securely store the content per license requirements. Not all data from vendors comes in a readily accessible structured form and may need to be reformatted.
- ECDS staff may provide basic instructions and assistance with TDM applications. If more intense assistance is required (e.g., use of programming languages such as Python), students or faculty should consider assembling a digital project proposal for ECDS. In many cases, the vendor will assist the patron directly in utilizing a walled garden or API. ECDS and/or subject librarians may also be able to assist.
- Emory Libraries will work closely with the ECDS) to provide access to TDM corpora.
- Emory Libraries and ECDS will work with online TDM tools and interfaces as they develop and explore possibilities for the inclusion of Emory's locally digitized content.
- In some cases, we can store and maintain physical media as necessary for accessing data. The Libraries may also ask for stand-alone purchased corpora in digital format to be locally preserved if feasible and if space allows.
- The Libraires do not guarantee storage space for these materials or high-speed computing resources for TDM or AI activities.
- The Libraries may point users to technical contacts at vendors to assist with downloading and/or delivery of data.